


# Sentence repetition with bilinguals with and without DLD: Differential effects of memory, vocabulary, and exposure

Amy S. Pratt<sup>1</sup> , Elizabeth D. Peña<sup>1</sup> and Lisa M. Bedore<sup>2</sup><sup>1</sup>University of California, Irvine and <sup>2</sup>Temple University

## Research Article

**Cite this article:** Pratt AS, Peña ED, Bedore LM (2021). Sentence repetition with bilinguals with and without DLD: Differential effects of memory, vocabulary, and exposure. *Bilingualism: Language and Cognition* **24**, 305–318. <https://doi.org/10.1017/S1366728920000498>

Received: 5 February 2019  
Revised: 3 July 2020  
Accepted: 6 July 2020  
First published online: 11 September 2020

**Keywords:**

bilingual development; language disorders; sentence repetition

**Address for correspondence:**

Amy S. Pratt, Email: [apratt1@uci.edu](mailto:apratt1@uci.edu)

**Abstract**

Though previous research has shown that sentence repetition (SR) is an informative tool for identifying developmental language disorder (DLD) in bilinguals, little is understood about the skills that underlie children's performance on the task. With a population of 136 school-age Spanish–English bilinguals, the present study explores the contribution of verbal short-term memory, vocabulary, and language exposure on two SR tasks developed in English and Spanish. Results indicate that these skills may differentially underlie SR in typical versus disordered populations. Whereas the strongest predictors of English SR performance for typical bilingual children were expressive vocabulary and language exposure, bilingual children with DLD relied most heavily on verbal short-term memory. ROC curves to determine the classification accuracy of SR were conducted. Classification accuracy was excellent, with area under the curve reaching .92 for the English SR task and .87 for Spanish SR.

**Introduction**

Data from the National Center for Education Statistics shows that nearly 8% of public school students in the U.S. – an estimated 3.7 million children – speak Spanish at home (McFarland, Husser, de Brey, Snyder, Wang, Wilkinson-Flicker, Gebrekristos, Zhang, Rathbun, Barmer, Bullock Mann & Hinz, 2017). The heterogeneity of children's language skills within this population is widely documented (Bohman, Bedore, Peña, Mendez-Perez & Gillam, 2010), as are the difficulties associated with identifying children with developmental language disorder (DLD) who are bilingual (Bedore & Peña, 2008). Sentence repetition (SR) tasks have emerged as a promising tool for the identification of bilingual children with DLD (Meir, Walters & Armon-Lotem, 2016; Thordardottir & Brandeker, 2013). Though different SR tasks vary with respect to length, complexity, and how they are scored, overall classification accuracy of bilinguals with DLD ranges from fair (>.80) to good (>.90) (Armon-Lotem & Meir, 2016; Fleckstein, Prévost, Tuller, Sizaret & Zebib, 2016). Despite this utility as a diagnostic instrument, the mechanisms that underlie children's performance on SR tasks are not fully understood. The objective of the current study, therefore, is the following: first, to replicate previous research that has evaluated the classification accuracy of SR tasks with bilinguals; and second, to investigate predictors of children's performance – specifically, their verbal short-term memory and lexical ability. Extant research suggests that SR relies on skills from each of these domains (Ebert, 2014; Simon-Cerejido & Méndez, 2018); however, the relative contribution of these factors, and their vulnerability to differing levels of language exposure, has yet to be systematically explored with bilinguals.

**Using SR to detect DLD**

DLD is characterized by significant difficulty in the area of oral language in children who do not otherwise present neurological, perceptual, or cognitive delays (Leonard, 2014; Stark & Tallal, 1981). SR tasks have been shown to correctly identify high percentages of children with DLD (sensitivity) without incorrectly flagging TD children (specificity) (Archibald & Joannis, 2009; Conti-Ramsden, Botting & Faragher, 2001). As a rule, tasks with sensitivity and specificity above .90 are considered “good” discriminators, while those with sensitivity and specificity between .80 and .90 are considered “fair” discriminators (Plante & Vance, 1994). Using a SR task taken from the Clinical Evaluation of Language Fundamentals, Third Edition (CELF-3; Semel, Wiig & Secord, 1995), Conti-Ramsden and colleagues (2001) detected DLD in 160 monolingual English speakers with sensitivity of .90 and specificity of .85, significantly better than both a nonword repetition task and a grammatical tense-marking task. Since then, a substantial body of work has corroborated the finding that English-speaking children with DLD perform more poorly on tasks of SR than TD children (Archibald & Joannis, 2009; Briscoe, Bishop & Frazier Norbury, 2001; Eadie, Fey, Douglas

& Parsons, 2002; Everitt, Hannaford & Conti-Ramsden, 2013; Redmond, 2005; Riches, 2012; Seeff-Gabriel, Chiat & Dodd, 2010).

Truly effective discriminators of impairment should extend to populations with diverse linguistic backgrounds. SR tasks have accurately differentiated children with DLD in non-mainstream dialects of English (Oetting, McDonald, Seidel & Hegarty, 2016) as well as in typologically diverse languages, such as Cantonese (Stokes, Wong, Fletcher & Leonard, 2006) and Czech (Smolík & Vávru, 2014). A growing number of studies has investigated the use of SR tasks with children acquiring two languages (Armon-Lotem & Meir, 2016; Ebert, 2014; Fleckstein et al., 2016; Gutiérrez-Clellen, Restrepo & Simón-Cerejido, 2006; Meir et al., 2016; Simón-Cerejido, 2017; Thordardottir & Brandeker, 2013; Tuller, Hamann, Chilla, Ferré, Morin, Prevost, Dos Santos, Abed Ibrahim & Zebib, 2018; Verhoeven, Steenge, van Weerdenburg & van Balkom, 2011; Ziethe, Eysholdt & Doellinger, 2013). Of these, two studies have used SR to detect impairment in Spanish-English bilinguals, with somewhat mixed results. A recent study by Simón-Cerejido (2017) used two SR tasks, one in English and one in Spanish, each with 21 sentences balanced for sentence length and sentence type across languages. Discriminant function analyses with 40 three-year-old bilinguals with DLD and 40 age-matched TD peers found that the Spanish SR task had good sensitivity (.93) and fair specificity (.80), whereas the English SR task had poor sensitivity (.59) but good specificity (.89). Working with slightly older preschoolers, Gutiérrez-Clellen et al. (2006) used a grammatical task that combined 23 Spanish morphosyntactic cloze targets and 51 Spanish SR targets. These items, chosen to maximize differences between DLD and TD groups, yielded sensitivity and specificity, each, of .86 for four-year-olds and .94 for five-year-olds. The present study builds on their original bank of SR items, which has been expanded to include sentences suitable for older school-age bilinguals.

### *Predictors of SR performance*

It is perhaps not surprising that children with DLD perform more poorly on SR tasks than their TD peers. What is less understood, however, are the mechanisms that drive these differences in performance. Experimental research into the nature of SR has concluded that it primarily taps children's grammatical knowledge, particularly their ability to comprehend and reconstruct syntactic structures (Kapantzoglou, Thompson, Gray & Restrepo, 2016; Kidd, Brandt, Lieven & Tomasello, 2007; Polišenská, Chiat & Roy 2015; Riches, Loucas, Baird, Charman & Simonoff, 2010), a language domain that is profoundly and characteristically difficult for children with DLD. Nonetheless, recent work suggests that additional skills and processes, beyond grammar, may influence children's ability to accurately produce the targeted grammatical constructions (Moll, Hulme, Nag & Snowling, 2015; Poll, Miller, Mainela-Arnold, Adams, Misra & Park, 2013; Riches, 2012). These studies describe two main types of predictors: those that implicate memory and those that implicate language. The former emphasizes the role of verbal working memory in linguistic tasks and posits that problems in verbal memory will compromise children's performance on SR (Alloway & Gathercole, 2005; Ebert, 2014). The latter, in contrast, emphasizes that the storage of information in verbal memory is critically dependent on the quality of one's linguistic representations and, of relevance to the present study, of one's lexical representations (Allen &

Hulme, 2006; Klem, Melby-Lervåg, Hagtvet, Lyster, Gustafsson & Hulme, 2015; Melby-Lervåg & Hulme, 2010).

Theoretical accounts that integrate memory and language, such as Baddeley's (2000, 2012) Multicomponent Model of Working Memory, are often invoked to make sense of the complex demands of SR tasks (e.g., Riches, 2012; Smolík & Vávru, 2014). Baddeley's model is comprised of a central control system of limited attentional capacity, termed the central executive, aided by two memory storage systems: (a) the phonological loop, responsible for maintaining verbal information and (b) the visuo-spatial sketchpad, responsible for maintaining visual information. A fourth component, the episodic buffer, was added to the model to act as an interface between the subsystems of working memory and long-term memory (LTM). Researchers stipulate that SR taps the capacity of the episodic buffer (Baddeley & Wilson, 2002), in particular, because the buffer integrates the information temporarily held in the phonological loop with the existing semantic and syntactic information held in LTM.

While Baddeley's framework is helpful for conceptualizing the linguistic and storage demands of SR, it does not account for bilingual-specific factors such as variation in first (L1) and second (L2) language knowledge and disparity in timing of L1 and L2 exposure (Genesee, Hamers, Lambert, Mononen, Seitz & Starck, 1978). Presumably, these factors will influence both a bilingual's working memory capacity (Thordardottir & Brandeker, 2013) and their long-term linguistic representations (Blom & Paradis, 2013), thus impacting their performance on SR tasks.

### *SR tasks and memory*

Several studies have shown that children's performance on SR tasks is related to their performance on other memory tasks (Alloway & Gathercole, 2005; Baddeley, Hitch & Allen, 2009; Ebert, 2014; Poll et al., 2013; Riches, 2012; Smolík & Vávru, 2014; Willis & Gathercole, 2001; Ziethe et al., 2013). Alloway and Gathercole (2005) reported that children who performed well on a battery of simple and complex memory tasks also performed well on SR. Simple tasks, such as digit span or nonword repetition, measure one's ability to temporarily hold in mind phonological material. Complex tasks, such as backward digit span or listening span tasks, involve manipulation of stored material. Our study focuses on the simple storage of phonological material, given that this is the component of verbal memory that is used in much of the research on SR to date.

A common way to quantify simple STM storage is with a nonword repetition (NWR) task (e.g., Ebert, 2014; Hesketh & Conti-Ramsden, 2013; Meir, 2017; Smolík & Vávru, 2014), which asks children to repeat a string of sounds that together form a nonce word, ensuring that the repetition of sounds depends on memory while minimizing reliance on existing lexical knowledge. However, a task that was once thought to be void of language experience is now considered to have the highest linguistic load of simple STM storage tasks (Meir, 2017). Children's NWR is related to prosodic structure, segmental complexity, and phonotactic probability (Chiat, 2015). Deficits in phonological STM may affect children's ability to develop stable long-term grammatical representations (Adams & Gathercole, 2000), which, in turn, may affect their accuracy when imitating sentences that contain those forms. According to early work in SR, children are only able to reproduce forms when they have some existing knowledge of that form (Carrow, 1974; Gallimore & Sharp, 1981).

Notably, STM storage appears to contribute more significantly to SR for children with impairment than it does for children with

TD. In a study of 11-year-old children with and without a reported history of DLD, Hesketh and Conti-Ramsden (2013) found that children's performance on NWR was significantly predictive of SR performance only for children with a history of DLD, but not for TD controls. In younger children, Riches (2012) reported that NWR was a strong and significant predictor for six-year-old children with DLD, but not for age- or language-matched controls. One plausible explanation of this work is that, because children with DLD have impaired language skills, they rely on memory to complete SR tasks.

The contribution of memory on SR with bilinguals with DLD is less settled. Ziethe et al. (2013) observed that SR scores for 15 four- and five-year-old German bilinguals with DLD were correlated with simple STM storage ( $r = .62, p = .07$ ), albeit marginally significant, as measured by a forward digit span task. However, the association between memory and SR for bilingual children with TD was weak and insignificant. With a sample of monolingual ( $n = 81$ ) and bilingual ( $n = 109$ ) Hebrew and Russian learners, Meir (2017) evaluated whether group differences between bilinguals with TD and specific language impairment persisted on SR, after controlling for measures of verbal STM. Results showed that the effect of impairment persisted in both languages, even after controlling for children's forward digit span and NWR scores, with comparable effect sizes across memory tasks. Meir (2017) concluded that it is poor linguistic representations, and not memory, that explain the SR performance of bilingual children with DLD.

### SR tasks and lexical knowledge

Considerable research has substantiated the notion that SR also taps children's lexical knowledge (Klem et al., 2015; Moll et al., 2015; Potter & Lombardi, 1990; Simon-Cerejido & Méndez, 2018; Stokes et al., 2006; Ziethe et al., 2013). Among bilinguals, there is evidence that a minimal amount of lexical knowledge may be necessary in order for bilinguals to be successful in SR in their weaker language (Simón-Cerejido, 2017). A recent study with Spanish-English bilinguals explored the relationship between lexical measures and children's SR performance. With 61 TD preschoolers, Simon-Cerejido and Méndez (2018) observed that a sizeable percentage of the variance in the SR scores in both English and Spanish was explained by children's scores on an expressive vocabulary test in the language tested (54% in English and 16% in Spanish). They attributed their findings to the large role that lexical knowledge plays in performance on SR, particularly among emerging bilinguals.

There is some, albeit limited, evidence that lexical knowledge contributes to the SR performance of bilinguals with DLD. A preliminary report of a SR task used with Spanish-English bilingual children reported that the number of individual words repeated correctly, as opposed to number of sentence structures repeated correctly, yielded the largest effect sizes between TD and DLD groups (Restrepo, Gorin, Gray, Morgan & Barona, 2010). An investigation of bilingual children with DLD acquiring German found a strong, zero-order association ( $r = .70, p < .05$ ) between children's expressive vocabulary scores and their performance on a German SR task (Ziethe et al., 2013); however, they did not report the strength of this association for TD bilingual controls. To our knowledge, no study has attempted to predict SR performance in bilinguals with DLD using both measures of verbal STM and lexical knowledge.

### SR tasks and language experience

Finally, bilingual children's performance on any language task is at least somewhat related to their experience in that language. Even among SR tasks, which are purported to be less affected by exposure than lexical tasks (Armon-Lotem, de Jong & Meir, 2015), children's experience in a language predicts their SR performance in that language (Thordardottir & Brandeker, 2013). What is particularly interesting about the effect of language exposure on SR is how it differentially relates to bilinguals with TD and those with DLD. Recent work with French bilinguals by Fleckstein et al. (2016) showed that language exposure was significantly and strongly correlated to performance on a SR task for children with TD ( $r = .48, p < .01$ ), but not for those with DLD.

These results support the notion that, in order for bilinguals to comprehend and repeat the linguistic forms that are featured by SR tasks, they must first be able to identify those forms from the language input. Bilinguals with DLD have inherent difficulty with this, even when their experience in a language is extensive. Bilinguals with TD are inherently good at this, even when their experience in a language is limited (see Armon-Lotem, 2017). What's more, language experience is shown to be related to the skills that underlie SR, including verbal STM (Thordardottir & Brandeker, 2013) and lexical knowledge (Anaya, Peña & Bedore, 2018).

### Present study

Previous research asserts that tasks like SR are "the most sensitive and specific tool for screening language impairment in bilingual children" (Armon-Lotem, 2017, p. 34). Much of this utility comes from the fact that SR taps children's grammatical abilities, a finding that is robust in the literature. However, a broader view of SR suggests that children's performance may also be related to their verbal STM, lexical knowledge, and language experience. Recent work provides rationale for investigating the relative contribution of these skills on bilinguals' SR performance concurrently. Meir and Armon-Lotem (2017) showed that bilingualism was associated with decreased vocabulary and lower performance on NWR and SR tasks among typical L2 Hebrew speakers; however, the negative effect of bilingualism on verbal STM disappeared once vocabulary was accounted for.

The extent to which these skills may differentially affect the SR performance of bilinguals with and without DLD remains an open question: some studies provide evidence that children with DLD recruit resources from verbal STM to complete SR tasks (e.g., Ziethe et al., 2013); others have concluded that simple memory tasks are insufficient to explain the differences between children with DLD and those with TD (Meir, 2017; Smolík & Vávru, 2014); still others have found differential effects related to language exposure (Fleckstein et al., 2016).

In addition to exploring these relationships, the present study assesses classification accuracy of two SR tasks, one in English and one in Spanish, with a population of school-age bilinguals with and without DLD. There is clinical relevance for evaluating the diagnostic accuracy of SR tasks in each language separately. In the U.S., where the current research was conducted, fewer than 10% of speech-language pathologists speak a language other than English (American Speech & Hearing Association, 2017). Therefore, to increase external validity, we compare diagnostic accuracy when SR tasks are administered in both languages, only in Spanish, and only in English. Specifically, we pose the following questions:



- 1) To what extent do verbal STM, lexical knowledge, and language experience explain variability in school-age Spanish-English bilinguals' performance on SR tasks in Spanish and English? Do these predictors differentially relate to the performance of children with DLD versus those with TD?
- 2) How accurately do English and Spanish SR tasks identify school-age Spanish-English bilinguals with DLD from their TD bilingual peers?

## Methods

### Participants

Participants' data comes from a longitudinal study of cross-linguistic outcomes of bilingual children with and without DLD (Bedore, Peña, Griffin & Hixon, 2016). Participants in the larger, longitudinal study ( $n = 361$ ) were recruited from preschool, first-, and third-grade classrooms in two public school districts in central Texas. The present analysis is comprised of a sub-sample of children who fit the following criteria at year one of the longitudinal portion of the study: children who were between the ages of 6;10 and 9;11 ( $n = 217$ ), had at least 20% exposure in each language ( $n = 160$ ), and had complete SR data in both languages. This yielded a total of 136 children: 26 children in the DLD group and 110 children in the TD group.

Children's language experience in English varied widely across our sample. To obtain information about English experience, research assistants interviewed parents and teachers using the Bilingual Input-Output Survey of the Bilingual English Spanish Assessment (BESA; Peña et al., 2018). To measure input, parents reported on the language their child was most likely to hear on an hour-by-hour basis during a typical weekday and a typical weekend day, and teachers reported on the language the child was most likely to hear on a half-hour basis during a typical school day. The daily hours spent in each language were then extrapolated to the remaining days of the week and summed. The total number of hours of input in English in a typical week was divided by the total number of hours of input overall (English plus Spanish), yielding a percentage of time the child spent hearing English in a typical week. To measure output, the process was repeated, but parents and teachers reported on the language the child was most likely to SPEAK. Because children's input and output percentages were highly correlated ( $r = .98, p < .001$ ), they were averaged to create a single variable, "English input/output." We excluded functionally monolingual children (i.e., those with less than 20% exposure to their other language, Bedore, Peña, Summers, Boerger, Resendiz, Greene, Bohman & Gillam, 2012) as these children lie on extreme ends of the spectrum of bilingualism (average English input/output in our sample for functionally monolingual English and Spanish speakers was 93% and 12%, respectively) and perform similarly to monolinguals.

Demographic information, including information about language experience, is displayed in Table 1. Independent samples  $t$ -tests showed that DLD and TD groups did not differ significantly on the demographic variables, except for English input/output  $t(134) = -2.57, p = .011$ . Although participants in both groups reported greater exposure to Spanish than to English – overall, the average English input/output was 45.65% ( $SD = 12.26\%$ ) and the average Spanish input/output was 54.35% ( $SD = 12.26\%$ ) – children with TD reported slightly higher (7%) English input/output than the children with DLD. Parents of children with TD and DLD reported that children, on average, were first exposed to English at approximately 2 years 10 months of age. Families'

socioeconomic status (SES) was computed using mother's education. The average Hollingshead score for our sample was 2.50 ( $SD = 1.56$ ), which corresponds to a junior high school (2) or partial high school (3) education (Hollingshead, 1975). Groups did not differ on SES, which has been shown to adversely affect both verbal STM and lexical knowledge in bilinguals (Meir & Armon-Lotem, 2017). Nearly all participants identified as Hispanic (99%). Twelve participants' SES responses were missing at random and nine participants were missing data on ethnicity.

### Procedures

Identification of DLD was carried out using a single-gate design in two phases, a screening phase and a confirmatory diagnostic phase, in order to reduce spectrum bias (Dollaghan & Horner, 2011). In the first phase, children were screened using the Bilingual English-Spanish Oral Screener (BESOS; Peña et al., 2010a), an experimental version of a semantics and morphosyntax screener with good classification accuracy (Lugo-Neiris, Peña, Bedore & Gillam, 2015). Children whose highest score fell below the 25th percentile on either the morphosyntax or the semantics subtest in their better language were considered to be at risk for DLD. Children at risk for DLD were oversampled (i.e., the prevalence of DLD in our sample exceeded the 7% in the population, Tomblin, Records, Buckwalter, Zhang, Smith & O'Brien, 1997) in order to reach adequate statistical power.

In the confirmatory diagnostic phase, DLD was confirmed if children met four of the following five indicators: (a) scored 1 SD below age norms on the BESOS screener in both languages, (b) scored 1 SD below age norms in both languages on the Bilingual English Spanish Assessment – Middle Extension (BESA-ME; Peña et al., 2010b) semantics subtest, (c) scored 1 SD below age norms in both languages on the BESA-ME morphosyntax subtest, (d) scored 1 SD below age norms on both the English version of the Test of Narrative Language (TNL; Gillam & Pearson, 2004) and an experimental version of the TNL adapted to Spanish (TNL-S; Gillam et al., in development), and (e) scored below an average of 4.25 (out of 5) in both languages on the Inventory to Assess Language Knowledge (ITALK) (Peña et al., 2018). This resulted in a DLD group of 26 children. We identified children with typical language skills if they scored above -1 SD of the mean on two or more of these measures in at least one language. Additionally, all participating children scored within normal limits on the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998) and passed an initial hearing screening administered by their school nurse.

Participants also completed assessments that examined their phonological STM and lexical knowledge. All testing occurred in a quiet space in children's schools in three or four sessions of 30 to 45 minutes. The order and language of testing was randomized across participants. All tests were administered by experienced research assistants following one-on-one training sessions with project staff, comprised of certified speech language pathologists who have worked with bilingual school-age children. Less than 10% of data for each variable of interest was missing by group, and missing data was deleted listwise.

### Identification measures

**BESOS.** The Bilingual English-Spanish Oral Screener (BESOS; Peña et al., 2010a) consists of two subtests, semantics and morphosyntax, in both English and Spanish. The semantics subtest uses receptive and expressive items to examine children's

**Table 1.** Descriptives of demographic, dependent, and independent variables, by group

	TD group <i>n</i> = 110			DLD group <i>n</i> = 26		
	Range	Mean	SD	Range	Mean	SD
<b>Demographics</b>						
Age (months)	81–138	103.45	12.33	86–126	100.19	10.88
SES <sup>a</sup>	1–6	2.61	1.59	1–6	2.39	1.50
Age of 1 <sup>st</sup> English exposure (years)	0–6	2.80	1.78	0–5	2.71	1.90
Gender	47% male			65% male		
Ethnicity	99% Hispanic			100% Hispanic		
<b>Dependent variables</b>						
English SR	49.87–122.80	93.42	19.58	49.87–102.91	62.37	15.27
Spanish SR	55.70–129.97	88.06	20.33	55.70–107.69	66.55	13.65
<b>Independent variables</b>						
English input / output	21.28–79.85	46.72	12.00	21.26–70.00	39.92	12.71
English NWR <sup>b</sup>	31–83	60.36	11.23	33–82	56.35	11.08
Spanish NWR <sup>c</sup>	49–100	80.19	11.92	36–93	69.23	12.81
English exp. vocabulary <sup>d</sup>	20–119	59.57	20.10	2–94	37.38	20.73
Spanish exp. vocabulary <sup>e</sup>	10–88	54.69	14.03	29–54	42.62	6.41

a = Hollingshead score (Hollingshead, 1973); b = Nonword repetition task (Dollaghan & Campbell, 1998); c = Experimental nonword repetition task in Spanish; d = Expressive One-Word Picture Vocabulary Test (EOWPVT; Brownell, 2000); e = Spanish Bilingual Edition of Expressive One-Word Picture Vocabulary Test (EOWPVT-SB; Brownell, 2001)

semantic breadth and depth, including functions, categories, definitions, characteristic properties, similarities and differences, and associations. The morphosyntax subtest measures grammatical constructions in each language that are challenging for children with DLD (e.g., past tense *-ed*, third person present tense *-s*, and copulas in English; articles, direct object clitics, and subjunctive in Spanish). Previous research using BESOS to predict DLD in first graders showed .95 sensitivity and .71 specificity when using a cut score of 1 SD below the mean in the child's best language (Lugo-Neris et al., 2015).

**BESA-ME.** A field test version of the Bilingual English Spanish Assessment-Middle Extension (BESA-ME; Peña et al., 2010b) was administered during the confirmatory diagnostic phase of testing to confirm DLD status. The semantics subtest measures expressive and receptive semantic knowledge by targeting children's knowledge of repeated associations, category generation, functions, definitions, similarities and differences, and analogies. The English subtest consists of 23 semantics items and the Spanish subtest consists of 26. The morphosyntax subtest targets morphosyntactic constructions known to be difficult for children with DLD in each language (e.g., English: singular present tense and past tense; Spanish: adjective agreement and direct object clitics). The subtest presents 18 grammatical cloze items for English and 19 cloze items for Spanish, as well as six sentences in each language to be repeated in a SR task. The sentences contain 22 and 26 targets in English and Spanish, respectively, which are scored as correct if the child included the target word in his or her repetition. Preliminary analyses of classification accuracy of the BESA-ME using a composite of the semantics score in the better language and the morphosyntax score in the better language indicate 1.00 sensitivity and .87 specificity for second graders, and 1.00 sensitivity and .95 specificity for fourth graders (Bedore et al., 2018).

**TNL / TNL-S.** The English Test of Narrative Language (TNL; Gillam & Pearson, 2004) and the Test of Narrative

Language - Spanish Experimental Version (TNL-S; Gillam et al., in development) assess children's narrative comprehension and production abilities. Though different stories are used for the English and Spanish versions, the structure of the test is parallel across languages. The test consists of six subtests in which the child is directed to retell a story he/she just heard, tell a story sequence based on a picture sequence, generate a story with a picture or respond to examiner questions. Standard scores ( $M = 100$ ,  $SD = 15$ ) were derived separately for the English and Spanish versions. For English, the TNL manual indicates a sensitivity of .92 and specificity of .87; sensitivity and specificity on the TNL-S showed sensitivity from .80 - .85 and specificity from .74 - .81 (Gillam et al., 2006).

**ITALK.** The Inventory to Assess Language Knowledge (ITALK; Peña et al., 2018) was used to measure parents' and teachers' perceptions of children's ability in each language. Parents and teachers rated children on a 5-point scale, from 1 (minimal proficiency) to 5 (high proficiency) in the following areas: vocabulary use, speech production (intelligibility), sentence production (utterance length), grammatical proficiency, and comprehension proficiency. The five scores in each language were averaged to yield a separate English and Spanish score based on parent and teacher reports. For cases in which a parent or teacher did not have knowledge of an area of one of the child's two languages, this was marked as "unknown" and was not included in the average (Gutiérrez-Clellen & Kreiter, 2003).

### Predictor measures

#### Nonword repetition

Children completed NWR tasks containing 16 items in each language that increased in syllable length. In English, nonwords ranged from one to four syllables and were developed by Dollaghan and Campbell (1998). In Spanish, nonwords ranged from two to

five syllables and were based on a list first developed by Calderón (2003). Both NWR tasks contained low word-like nonwords for each language. The English list excluded later-developing sounds, consonant clusters, and tense vowels that occurred more than once in a word. The Spanish list excluded later-developing sounds, as well as syllables that occurred more than 200 times in the Alameda and Cuetos (1995) corpus, and it included only tense vowels. Previous work comparing both lists of nonwords showed that, because of the simpler CVCV phonological structure of Spanish, the difficulty across the English and Spanish tasks is not equivalent in bilinguals (Summers, Bohman, Gillam, Peña & Bedore, 2010). Therefore, the present study added five syllable nonwords to the original Calderón (2003) items, in order to make the task difficulty more comparable across languages (the final list of Spanish NWR items is available in online supplementary materials, Supplementary Material).

A native bilingual adult male speaker digitally recorded nonwords in both languages. Stimuli were presented on a laptop computer with the use of headphones and instructions were provided in the language of the stimuli. Responses were recorded using a microphone and digital recorder. These recordings were then transferred to a computer and transcribed in English and in Spanish by bilingual research assistants. Transcriptions were scored for percentage of phonemes correct (PPC) following the protocol in Dollaghan and Campbell (1998). A consonant was scored as incorrect if it was omitted or deleted. Participants were not penalized for distortions. As part of data checking procedures, 10% of the NWR responses for the total sample was independently re-transcribed and scored; overall reliability at the phoneme level across languages was 83%, which is comparable to previous work reporting reliability between 83–85%, when utilizing the same strict criteria for agreement on voice, place, and manner of articulation (Krishnan, Alcock, Mercure, Leech, Barker, Karmiloff-Smith & Dick, 2013; Topbaş, Kaçar-Kütükçü & Kopkalli-Yavuz, 2014). Three children with TD were missing English and Spanish NWR scores, at random, in our sample and were removed listwise from regression analyses.

### *Expressive vocabulary*

The Expressive One-Word Picture Vocabulary Test (EOWPVT; Brownell, 2000) and Spanish Bilingual Edition (EOWPVT-SB; Brownell, 2001) were administered to assess lexical knowledge in English and Spanish, respectively. These are norm-referenced tests of single-word expressive vocabulary each consisting of 170 items that follow a developmental sequence. The EOWPVT and EOWPVT-SB present examinees with colored line drawings that depict an action, object, category, or concept, and examinees must label each drawing. Whereas the English EOWPVT was administered and scored according to the manual, the SB version was administered and scored only in Spanish, to reflect children's language-specific knowledge in Spanish. Basal and ceiling rules provided in the test manuals were used to compute raw scores. 11 children with TD were missing EOWPVT data and 16 were missing EOWPVT-SB data. They were not included in our examination of the first research question.

### *Sentence repetition measures*

SR items target grammatical constructions that are challenging for children with DLD but are not easily elicited using a cloze task. SR items developed during pilot work for the Bilingual English Spanish Assessment (BESA; Peña et al., 2018) – a language

assessment tool for bilingual children ages four through six – formed the basis for the present study. In expanding the SR items for a school-age population, sentences from the BESA dataset that continued to show growth and reliably differentiate older TD and DLD children were retained (6 sentences in English and 4 sentences in Spanish) and new sentences were added (5 sentences in English and 6 sentences in Spanish). Examples of sentences that were retained include, in English, “The teacher wants to know who brought the snake” and, in Spanish, “La señora llamó a los bomberos cuando vio que salía humo del carro” [translation: The woman called the firemen when she saw that smoke was coming out of the car]. Both sentences are multiclausal and require children to mark verbs using various forms of tense and aspect. Examples of sentences that were added for use with school-age children include “Can she find a tank big enough for those fish?” and “Si tuviera un caballo, lo montaría todas las mañanas” [translation: If I had a horse, I would ride it every morning]. The new sentences in English SR included *do* insertion for negatives and questions with auxiliary; new sentences in Spanish included negative concord and subjunctive. The final SR task comprised 11 sentences in English and 10 sentences in Spanish. Given that sentences were selected based on their discriminant values (<.30), the final sentences differ across languages in terms of length and syntactic constructions. Sentences in English ranged in length from 6 to 11 words (average sentence: 8.91 words; 10.36 syllables). Sentences in Spanish ranged in length from 9 to 17 words (average sentence: 11.30 words; 20.80 syllables). Sentences for the English and Spanish tasks are described in detail in online supplementary materials (Supplementary Material).

SR tasks can be scored in various ways, depending on whether they are used for research or clinical purposes. In order to simulate clinical application, our SR items were scored during testing by research assistants using a dichotomous scoring scheme, similar to the scoring scheme prescribed by the Test of Language Development – Primary (TOLD-P-4; Newcomer & Hammill, 2008). Children received a score of 1 for each sentence repeated verbatim, with no additions, substitutions, or omissions, and a score of 0 for each sentence that deviated from the target sentence. Dichotomous scores were summed, for a maximum raw score of 11 in English and 10 in Spanish. Using a normative dataset of 614 bilingual children who received the same items, we derived standard scores based on the mean and SD for each age group (using one-year intervals). Standard scores had a mean of 100 and SD of 15. Bilinguals with balanced ability or dominance in English ( $n = 528$ ) were included in the English SR norm; those with balanced ability or dominance in Spanish ( $n = 343$ ) were included in the Spanish SR norm.

### *Analytical strategy*

Prior to exploring concurrent relationships between variables and SR, we conducted zero-order correlations to explore the bivariate relationships between SR scores in both languages and language-specific NWR, expressive vocabulary, and percentage English input/output, by group. The Benjamini–Hochberg procedure was used to control the rate of false discovery (Benjamini & Hochberg, 1995). Using this procedure, at a false discovery rate of 0.05, a correlation was considered significant when its  $p$  value was 0.016 or less. Next, hierarchical linear regressions were conducted as a follow-up analysis to the correlation analysis, as it allows us to assess the unique contribution of each independent variable to SR, relative to other predictors. After controlling

for age, we entered English input/output, NWR in the language of testing, and expressive vocabulary in the language of testing into the regression, to find the combination of variables most highly associated with SR scores in each language. Finally, we ran receiver operating characteristic (ROC) curves with children's SR scores in English, Spanish, and the score from the language in which they scored highest. Sensitivity, specificity, and overall classification accuracy are reported for each task.

## Results

Preliminary data analyses showed that our data were normally distributed. As seen in Table 1, on average, children with DLD scored lower than TD peers on all tasks. With respect to SR, the average child with DLD scored 59.21 ( $SD = 12.22$ ) on the English SR task and 63.83 ( $SD = 8.76$ ) on the Spanish SR task, which have a standardized mean of 100. In contrast, the average child with TD scored nearly 30 points higher in each language. Similar trends were observed with both NWR and expressive vocabulary, as children with DLD scored approximately one standard deviation below controls.

### Research aim 1: Predictors of SR

#### Bivariate relationships

The first research question aimed to characterize the predictors of SR for school-age bilinguals. We first conducted zero-order correlations between SR and all independent variables, by group. The correlations for children in the TD and DLD groups are shown in Table 2. Age was significantly related to SR performance in both groups, with the exception of Spanish SR among TD children. Among children in the TD group, English SR was significantly correlated with all language-specific IVs, with English expressive vocabulary showing the strongest relationship to SR ( $r = .62$ ,  $p < .001$ ), follow by English input/output ( $r = .47$ ,  $p < .001$ ), and finally English NWR ( $r = .27$ ,  $p < .05$ ). Similar trends were observed in Spanish in the TD group, with the exception of English input/output, which did not correlate with TD children's SR performance in Spanish. Among children in the DLD group, the bivariate correlations with SR in both languages showed distinct patterns. Unlike their TD counterparts, English expressive vocabulary ( $r = .24$ ,  $p = .228$ ) was unrelated to the English SR performance of bilinguals with DLD and English input/output was just significant, with a Benjamini-Hochberg adjusted  $p$ -value of .050 ( $r = .43$ ). However, the correlation between English NWR and English SR was strong and significant ( $r = .47$ ,  $p < .05$ ). Figure 1 features a scatterplot of this data. The non-associations between SR and expressive vocabulary in the DLD group are evidenced by flat black lines in the scatterplots; in contrast, the sloped dashed lines signify moderate-to-strong relationships between SR and language-specific IVs for children with TD. Finally, in addition to the bivariate correlations between predictors and SR, we also ran correlations to test for significant collinearity between age and English input/output. Age and English input/output significantly correlated among children with TD ( $r = .42$ ,  $p < .001$ ), but no relationship between age and exposure was found among children with DLD ( $r = -.12$ ,  $p = .362$ ).

#### Hierarchical regressions

Given the significant bivariate relationships observed in Figure 1, our next step was to describe the relative contribution of each predictor to SR when examined concurrently. We conducted

**Table 2.** Bivariate correlations between SR scores and predictor variables, by group

	English SR		Spanish SR	
	TD group	DLD group	TD group	DLD group
Age	$r = .34^*$	$.55^*$	.02	$.49^*$
English input/output	$.47^*$	$.43^*$	-.13	.13
English NWR	$.27^*$	$.47^*$	.02	.01
English exp. vocabulary	$.62^*$	.24	.03	-.25
Spanish NWR	$.29^*$	.10	$.24^*$	.10
Spanish exp. vocabulary	.02	-.17	$.52^*$	$.45^*$

\* = Correlation is significant after controlling for false discovery rate at 0.05 (Benjamini & Hochberg, 1995).

hierarchical regressions for each ability group, using forced entry of the following predictors in four blocks: (1) age, (2) English input/output, (3) NWR in the language of assessment, and (4) expressive vocabulary in the language of assessment. Because previous research has found that children's performance on SR is language specific, we did not test for cross-linguistic effects (Simon-Cerejido & Méndez, 2018). SR scores in English and Spanish were entered, in separate regressions, as the dependent variables. Upon running the regressions, we tested for multicollinearity using variance inflation factors (VIF) and tolerance. When VIF values exceed 4.0 or tolerance levels are less than .20, there may be a problem with multicollinearity (Hair, Black, Babin & Anderson, 2014). For our data, all VIF values were less than 1.49 and all tolerance values were greater than .66.

Predictor variables were added one block at a time in order to evaluate the unique contribution of each additional predictor to the model. In all four models, age was entered initially, at block 1, to control for developmental effects in all subsequent blocks. The English SR results with the TD group are shown in the upper part of Table 3. At block 1, age accounted for 12% (adjusted  $R^2 = .11$ ) of the variation in children's English SR scores,  $F(1, 95) = 12.53$ ,  $p < .01$ . At block 2, English input/output was added to the model and significantly improved the variance accounted for to 26%,  $F(1, 94) = 17.43$ ,  $p < .001$ , which made age insignificant. Part correlations for age and exposure, respectively, totaled .13 and .37, indicating that less than 2% of variance in English SR scores is uniquely explained by age. At block 3, English NWR was added to the model, which did not significantly improve the model,  $F(1, 93) = 1.99$ ,  $p = .162$ , as it accounted for just .02% of additional variance. Finally, English expressive vocabulary was added at block 4 and yielded a significant improvement in the overall proportion of variance accounted for,  $F(1, 92) = 26.60$ ,  $p < .001$ , explaining an additional 16% of the variance in children's SR abilities in English in the TD group. After 4 blocks, the resulting model accounted for approximately 43% of variability in children's English SR scores, with English input/output and English expressive vocabulary retaining individual significance.

Given the group differences observed in Figure 1, there is reason to suspect that the relationship between SR and the predictor variables may change relative to children's ability. For instance,



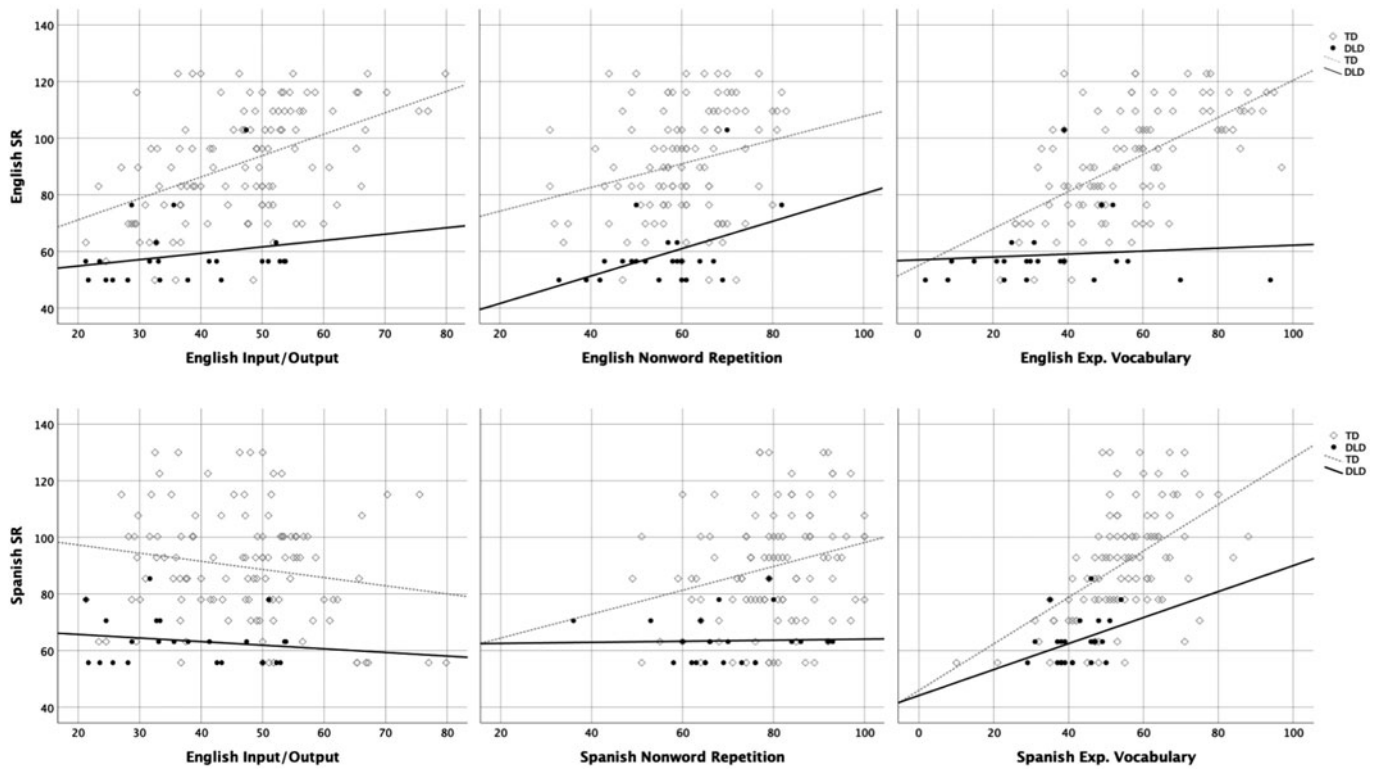


Fig. 1. Scatterplots of SR and language-specific predictors, by group

Table 3. Model summaries of regressions predicting English SR scores in TD and DLD groups

	Block 1		Block 2		Block 3		Block 4	
	$\beta$	t	$\beta$	t	$\beta$	t	$\beta$	t
TD group								
Age	.34	3.54**	.19	1.97	.18	1.85	-0.01	-0.05
English input/output			.40	4.18**	.37	3.72**	.22	2.43*
English NWR					.13	1.41	.08	.96
English exp. vocabulary							.50	5.16**
R <sup>2</sup>		.12		.26		.27		.43
R <sup>2</sup> change		.12**		.14**		.01		.16**
DLD group								
Age	.55	3.25**	.48	2.88*	.44	2.85*	.53	3.09**
English input/output			.32	1.94	.27	1.74	.27	1.74
English NWR					.35	2.34*	.42	2.62*
English exp. vocabulary							.21	-1.18
R <sup>2</sup>		.31		.40		.52		.52
R <sup>2</sup> change		.31**		.10		.12*		.00

\* =  $p < .05$ ; \*\* =  $p < .01$

bivariate correlations showed that English expressive vocabulary was strongly related to English SR among TD children ( $r = .47$ ,  $p < .001$ ) but was not related to English SR for children with DLD ( $r = -.13$ ,  $p = .935$ ). Hence, we ran an identical hierarchical regression with children with DLD, predicting their English SR

scores (see bottom half of Table 3). Results contrasted with what was found in the TD group. At block 1, age accounted for 31% (adjusted  $R^2 = .28$ ) of the variation in children’s English SR scores,  $F(1, 95) = 10.54$ ,  $p < .01$ . At block 2, English input/output was added to the model but did not significantly improve the



**Table 4.** Model summaries of regressions predicting Spanish SR scores in TD and DLD groups

	Block 1		Block 2		Block 3		Block 4	
	$\beta$	t	$\beta$	t	$\beta$	t	$\beta$	t
TD group								
Age	.02	.23	.11	0.94	.08	0.73	-.12	-1.13
English input/output			-.18	-1.55	-.23	-1.96	-.05	-0.50
Spanish NWR					.27	2.60*	.19	2.10*
Spanish exp. vocabulary							.52	5.43**
R <sup>2</sup>		.00		.03		.09		.32
R <sup>2</sup> change		.00		.03		.07*		.23**
DLD group								
Age	.50	2.75*	.49	2.60*	.51	2.67*	.33	1.74*
English input/output			.01	0.06	-.02	-0.11	.24	1.15
Spanish NWR					.15	0.80	.07	0.38
Spanish exp. vocabulary							.47	2.31*
R <sup>2</sup>		.24		.24		.26		.41
R <sup>2</sup> change		.24*		.00		.02		.15*

\* =  $p < .05$ ; \*\* =  $p < .01$

variability accounted for,  $F(1, 23) = 3.76$ ,  $p = .065$ , though age remained significant. At block 3, English NWR was added to the model and increased the  $R^2$  to 52% (adjusted  $R^2 = .46$ ),  $F(1, 22) = 5.49$ ,  $p = .029$ . Finally, the addition of English expressive vocabulary at block 4 did not result in significant improvement in the overall proportion of variance accounted for,  $F(1, 21) = 26.60$ ,  $p = .253$ . Overall, the final model accounted for nearly half of the variability in SR scores of children with DLD, with age and NWR retaining individual significance.

The Spanish results for children with TD and DLD are shown in Table 4. Among children with TD, neither age nor English input/output explained variability in Spanish SR scores (adjusted  $R^2 = .02$ ). However, at block 3, Spanish NWR accounted for 7% additional variability,  $F(1, 90) = 6.77$ ,  $p = .01$ . At block 4, Spanish expressive vocabulary accounted for an additional 23% of variability,  $F(1, 89) = 29.51$ ,  $p < .001$ , resulting in a model that explained 32% of variance in Spanish SR scores (adjusted  $R^2 = .29$ ).

Among children with DLD, age at block 1 was a significant predictor of Spanish SR  $F(1, 24) = 7.54$ ,  $p = .01$ , accounting for nearly 24% of variability (adjusted  $R^2 = .21$ ). At block 2, English input/output did not contribute significantly to the model,  $R^2$  change = .00. Likewise, at block 3, the addition of Spanish NWR was not a significant predictor of SR scores in Spanish, contributing only 2% of additional variance. At block 4, Spanish expressive vocabulary accounted for an additional 15% of variability,  $F(1, 21) = 5.32$ ,  $p = .031$ , resulting in a final model that explained 41% of variance in Spanish SR scores (adjusted  $R^2 = .30$ ).

### Research aim 2: Classification accuracy

The second research question sought to assess the classification accuracy of our SR tasks. In addition to English and Spanish, we included a third metric “best language,” determined by comparing each child’s performance in English and Spanish and inserting whichever score was higher. Receiver operating characteristic (ROC) curves were used to estimate cut scores that

optimize sensitivity and specificity in English, Spanish, and best language. The optimal cut scores denote the probability threshold for classifying a child as having DLD. We then estimated the positive and negative likelihood ratios for each language using the optimal cut scores for each task. The positive likelihood ratio corresponds to the ratio of the probability of correctly classifying a child as having DLD and the probability of incorrectly classifying a child as having DLD (i.e., sensitivity / 1-specificity). The negative likelihood ratio corresponds to the ratio of the probability of incorrectly classifying a child as TD and the probability of correctly classifying a child as TD (i.e., 1-sensitivity / specificity) (McGee, 2002). These results are reported in Table 5.

Additionally, the area under the curve (AUC) was calculated to determine the classification accuracy of each SR task. Conceptually, the AUC is the probability that the SR task will rank a randomly chosen child with TD higher than a child with DLD. Typical benchmarks of AUC conclude that an AUC between .70 and .80 is considered an acceptable discriminator; an AUC between .80 and .90 is considered an excellent discriminator; and an AUC above .90 is considered an outstanding discriminator (Rice & Harris, 2005).

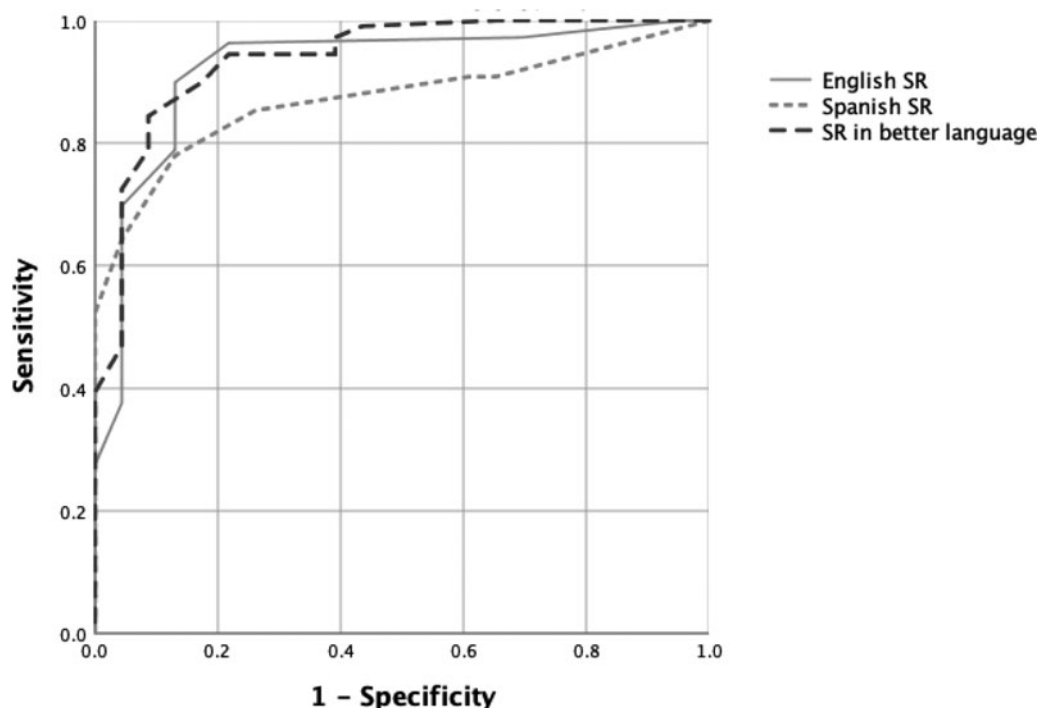
The three ROC curves are shown in Figure 2. The AUCs for English, Spanish, and best language are equal to .92, .87, and .94, respectively, which indicate that SR is an excellent discriminator of impairment in Spanish and an outstanding discriminator of impairment in English and in a child’s best language. Sensitivity and specificity are also reported in Table 5. Across all tasks, sensitivity estimates ranged from .86 to .91 and, specificity estimates, from .78 to .90.

### Discussion

SR tasks have been shown to be an informative tool in discriminating children with and without DLD in both monolingual (e.g., Conti-Ramsden et al., 2001; Seeff-Gabriel et al., 2010) and

**Table 5.** Results of ROC curves for SR tasks

	AUC	Std. Error	Asymptotic 95% CI		Optimal Cut	Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio
English SR	.92	.04	.85	.99	66.45	.86	.90	4.10	.15
Spanish SR	.87	.03	.81	.93	74.27	.86	.78	3.93	.28
SR best language	.94	.03	.88	1.00	80.50	.91	.86	5.54	.11

**Fig. 2.** ROC curves of SR tasks

bilingual populations (e.g., Meir et al., 2016; Tuller et al., 2018; Ziethe et al., 2013). Given the complex task demands inherent in SR, it is not unexpected that children with DLD perform consistently below TD controls on SR. However, pinpointing what contributes to their underperformance is more challenging. This is especially true for young bilinguals whose language experiences can vary greatly. Thus, the purpose of this study was twofold: first, to explore the memory and linguistic mechanisms that underlie SR for both children with DLD and TD; and, second, to evaluate the classification accuracy of SR tasks administered in English and Spanish with school-age bilingual children.

With respect to the first aim, extant literature shows that SR tasks require children to utilize both memory and language to varying degrees (Klem et al., 2015; Riches, 2012), and that language experience plays a role in the SR performance of bilingual children (Fleckstein et al., 2016). Our results both support these previous findings and add nuance to them. Concurrently examining relationships between SR, memory, lexical knowledge, and experience in a hierarchical linear regression allowed us to disentangle the unique contribution of each predictor. Several relationships that were statistically significant in a correlation analysis were no longer significant once we controlled for developmental and/or exposure effects. For example, among TD children, the bivariate correlation between English NWR and English SR was

moderate and statistically significant ( $r = .27$ ). However, after controlling for age and English input/output in our regression model, the contribution of NWR on English SR for TD children explained less than 1% of additional variance and was no longer a significant predictor of children's performance. This suggests, as previous research has found (Gibson, Summers, Peña, Bedore, Gillam & Bohman, 2015; Summers et al., 2010), that language experience plays a large role in the efficiency of phonological STM among TD bilinguals, and that the relationship between NWR and SR is at least partially explained by that experience.

We also found evidence that memory differentially contributes to SR scores for children with TD versus those with DLD. Among our group with DLD, we observed a significant bivariate association between NWR in English and SR in English ( $r = .47$ ), which replicated previous work by Ebert (2014), who reported an association of comparable strength ( $r = .38$ ) among a group of SE bilinguals with DLD similar in age to our participants. In contrast to what we found among TD children, the significant relationship between English NWR and English SR was maintained for children with DLD in the subsequent regression analysis. NWR accounted for 12% of additional variance in SR scores among the DLD group, after controlling for age and English input/output. These findings suggest that bilinguals

with DLD recruit resources from phonological STM to complete SR in their L2 English, whereas bilinguals with TD do not.

Turning to linguistic predictors, our results also support the view that bilinguals with TD rely on lexical knowledge to support SR performance. We found a strong association between language-specific expressive vocabulary and SR tasks in our TD group. Expressive vocabulary in English and Spanish explained an additional 16% and 23% of variability in English and Spanish SR scores of TD children, respectively, after controlling for age, exposure, and phonological STM. However, as was true of NWR, it appears that lexical knowledge differentially affects bilinguals with TD and DLD. Whereas expressive vocabulary was a strong predictor of children's SR in both English and Spanish among bilinguals with TD, for the DLD group this result was only significant in Spanish (accounting for 15% of variability). For children with DLD, English expressive vocabulary explained no additional variance in English SR.

Indeed, this underscores an important finding: we observed striking variation in the relative influence of certain predictors on English SR scores across TD and DLD groups. Among children with TD, the strongest predictor of performance was English expressive vocabulary. Among children with DLD, the strongest predictor was English NWR. This finding suggests that limited L2 vocabulary knowledge may be a source of difficulty for bilinguals with TD on English SR tasks; however, as children acquire more vocabulary, English SR scores improve. In contrast, greater L2 vocabulary knowledge among children with DLD is not related to better performance on SR. Perhaps this is due to the heterogeneity of vocabulary knowledge in bilingual children with DLD, on the one hand, and to the difficulties with complex syntax that are more apparent in older children with DLD, on the other. The complex syntactic knowledge tapped by SR tasks (e.g., Polišenská et al., 2015) does not appear to correlate linearly with L2 lexical knowledge among school-age children with DLD.

Differential trends were also observed across groups for language exposure and English SR. Regression results showed distinct effects of exposure by group, such that exposure was a significant predictor of English SR in the final model among children with TD ( $t = 2.43$ ,  $p = .017$ ) but was not significant in the final model among children with DLD ( $t = 1.74$ ,  $p = .096$ ). This result replicates findings reported by Fleckstein et al. (2016), who also showed a significant bivariate relationship between exposure and SR for French bilinguals with TD but not those with DLD. Their results, and ours, reaffirm that the language-learning difficulties of bilinguals with DLD are neuro-developmental in nature, and are not a result of low L2 exposure.

We observed patterns with respect to predictors of children's SR performance across their two languages. For instance, in Spanish – the first language of children in our sample – the strongest predictor of Spanish SR was Spanish expressive vocabulary for children in both the TD and DLD groups, accounting for 23% and 15% of variability, respectively. Children with TD also recruited some resources from phonological STM in Spanish, as evidenced by the significant contribution of Spanish NWR. However, Spanish NWR was not significantly related to Spanish SR scores for children with DLD. This is particularly interesting, given that the sentences that comprise our Spanish SR task were, on average, longer than the sentences comprising the English SR task. If differing sentence lengths were to differentially recruit memory resources, we would have expected the Spanish task to relate more strongly with memory. With respect to English –

the second language of children in our sample – regressions showed that expressive vocabulary and English input/output significantly contributed to SR among TD bilinguals; however, unlike in their L1, NWR was not significant. Among children with DLD, English SR performance was not predicted by expressive vocabulary, as in their L1, but by age and NWR. These differences across groups and languages may reveal differences in how children with DLD versus children with TD approach SR tasks, particularly in their second language. Whereas children with TD appeared to rely most heavily on the activation of recent lexical items, children with DLD, lacking lexical breadth in English, may have been more dependent on phonological STM to support SR in English.

The second research aim examined classification accuracy of SR tasks in English and Spanish. Given that SR requires the integration of skills from both memory and language, it has been shown to be a promising discriminator of impairment with bilinguals. Our results indicated AUCs of .92 and .87 for English and Spanish, respectively, indicating excellent discrimination capacity. The AUC improved to .94 when we utilized the SR score from the language in which each child scored the highest. The high sensitivity and specificity suggest that SR tasks could serve as an effective screener of language ability. This is particularly true when SR tasks are designed to contain grammatical targets that are typically problematic for children with DLD, thus maximizing differences between DLD and TD children, as these were. Although sensitivity on the SR task in children's best language was particularly high (.90), the English SR task by itself yielded acceptable levels of sensitivity (.86), which would apply to the English-only contexts in U.S. practice.

In sum, this work contributes to our theoretical understanding of DLD in bilinguals and, more specifically, the mechanisms that underlie SR performance in this population. Our results suggest that SR tasks differentially tap bilinguals' memory and lexical knowledge, depending on the language ability status of the child and the language of the task. Previous research has posited that SR tasks measure children's capacity to retrieve and apply grammatical representations from LTM (e.g., Moll et al., 2015; Poll et al., 2013; Riches, 2012). Indeed, our results confirmed that, among children with TD and children with DLD, L1 grammatical representations, as measured by SR, were predicted most strongly by L1 lexical knowledge. Because bilinguals are still in the process of acquiring and stabilizing their L2 grammatical representations in LTM, it is possible that the mechanisms underlying SR performance in the L2 differ from the L1. In fact, among children with TD, L2 lexical knowledge was the strongest predictor of L2 SR, followed by their L2 experience. However, for children with DLD who present with severe difficulties in L2 syntax, L2 lexical knowledge did not predict L2 SR performance. Instead, their performance on SR was predicted most strongly by L2 verbal STM. Importantly, this performance was unrelated to their language experience, thus confirming the notion that DLD is not attributable to lack of experience in a language: rather, it reflects an innate difficulty with language-learning.

Implications for practice can be drawn from the first and second aims. Regarding the former, we observed that neither English exposure nor English expressive vocabulary were significantly related to the performance of children with DLD on SR tasks in English after controlling for age, even though both of these measures were significantly related to the performance of children with TD. This result suggests that "quick fixes," like increasing children's exposure to English or improving their

lexical naming, may not be sufficient to resolve their difficulties with complex tasks requiring both memory and language. Instead, integrative intervention approaches that embed vocabulary targets into larger noun and verb phrases are recommended (Bedore, Peña, Fiestas & Lugo-Neris, 2020). Regarding the second aim, our results present a very efficient measure for screening school-age bilinguals for DLD – one that can be applied in under ten minutes and does not require laborious transcription.

The efficiency of our SR measure is also a limitation. We opted to use a dichotomous scoring scheme, which has high social validity, though it prohibited us from conducting an error analysis of challenging targets within each sentence. As recommended by Armon-Lotem and Meir (2017), and others, future research into SR should analyze these error patterns, taking note of how performance on specific targets may change with varying levels of language dominance. Given our limited power, we limited our analyses to within-language predictors. Future research should explore cross-linguistics of SR. Finally, while our results suggest differential effects of phonological STM and lexical knowledge, future research should further explore differences in how bilinguals with DLD and TD approach SR tasks using a variety of predictors. Because NWR is the most linguistically loaded of simple STM tasks, it is difficult to determine whether the significant contribution of NWR stems from the linguistic nature of the task. Until there is a body of research utilizing diverse tasks that vary in design, structure, and items, we cannot be certain the extent to which the differences in performance in this study are due to our particular tasks.

**Supplementary Material.** For supplementary material accompanying this paper, visit <https://doi.org/10.1017/S1366728920000498>

List of supplementary materials:

- S1 – Spanish Nonword Repetition Items
- S2 – Description of English Sentence Repetition Items
- S3 – Description of Spanish Sentence Repetition Items

**Acknowledgements.** This work was supported by National Institute on Deafness and Other Communication Disorders (NIDCD) Grant R01DC010366, PI: Peña. This report does not necessarily reflect the views or policy of the NIDCD. We thank the many families, research associates, and research assistants who contributed to this work.

## References

- Adams AM and Gathercole SE (2000) Limitations in working memory: Implications for language development. *International Journal of Language & Communication Disorders*, **35**, 95–116.
- Alameda JR and Cuetos Vega F (1995) *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Oviedo: Universidad de Oviedo.
- Allen R and Hulme C (2006) Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, **55**, 64–88.
- Alloway TP and Gathercole SE (2005) The role of sentence recall in reading and language skills of children with learning difficulties. *Learning and Individual Differences*, **15**, 271–282.
- American Speech & Hearing Association. (2017) Self-identify as a bilingual service provider. *ASHA Leader*, **22**, 61.
- Anaya JB, Peña ED and Bedore LM (2018) Conceptual scoring and classification accuracy of vocabulary testing in bilingual children. *Language, Speech, and Hearing Services in Schools*, **49**, 85–97.
- Archibald LMD and Joanisse MF (2009) On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research*, **52**, 899–914.
- Armon-Lotem S (2017) Disentangling bilingualism from SLI: dissociating exposure and input. *Bilingualism: Language and Cognition*, **20**, 33–34.
- Armon-Lotem S, de Jong J and Meir N (2015) *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Bristol: Multilingual Matters.
- Armon-Lotem S and Meir N (2016) Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: Evidence from Russian and Hebrew. *International Journal of Language & Communication Disorders*, **51**, 715–731.
- Baddeley AD (2000) The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, **4**, 417–423.
- Baddeley AD, Hitch GJ and Allen RJ (2009) Working memory and binding in sentence recall. *Journal of Memory and Language*, **61**, 438–456.
- Baddeley AD (2012) Working memory: Theories, models, and controversies. *Annual Review of Psychology*, **63**, 1–29.
- Baddeley AD and Wilson BA (2002) Prose recall and amnesia: Implications for the structure of working memory. *Neuropsychologia*, **40**, 1737–1743.
- Bedore LM and Peña ED (2008) Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism*, **11**, 1–29.
- Bedore LM, Peña ED, Anaya JB, Nieto R, Lugo-Neris MJ and Baron A (2018) Understanding disorder within variation: Production of English grammatical forms by English language learners. *Language, Speech, and Hearing Services in Schools*, **49**, 277–291.
- Bedore LM, Peña ED, Fiestas C and Lugo-Neris MJ (2020) Language and literacy together: Supporting grammatical development in dual language learners with risk for language and learning difficulties. *Language, Speech, and Hearing Services in Schools*, **51**, 282–297.
- Bedore LM, Peña ED, Griffin ZM and Hixon JG (2016) Effects of age of English exposure, current input/output, and grade on bilingual language performance. *Journal of Child Language*, **43**, 687–706.
- Bedore LM, Peña ED, Summers CL, Boerger KM, Resendiz MD, Greene K, Bohman TM and Gillam RB (2012) The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition*, **15**, 616–629.
- Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.
- Blom E and Paradis J (2013) Past tense production by English second language learners with and without language impairment. *Journal of Speech, Language, and Hearing Research*, **56**, 1–14.
- Bohman TM, Bedore LM, Peña ED, Mendez-Perez A and Gillam RB (2010) What you hear and what you say: Language performance in Spanish English bilinguals. *International Journal of Bilingual Education and Bilingualism*, **13**, 325–344.
- Bracken BA and McCallum RS (1998) *Universal Nonverbal Intelligence Test*. Chicago: Riverside Publishing Company.
- Briscoe J, Bishop DVM and Frazier Norbury C (2001) Phonological processing, language, and literacy: A comparison of children with mild-to-moderate sensorineural hearing loss and those with Specific Language Impairment. *Journal of Child Psychology and Psychiatry*, **42**, 329–340.
- Brownell R (2000) *Expressive One-Word Picture Vocabulary Test*. Novato: Academic Therapy.
- Brownell R (2001) *Expressive One-Word Picture Vocabulary Test – Spanish Bilingual*. Novato: Academic Therapy.
- Calderón J (2003) *Working memory in Spanish-English bilinguals with language impairment*. PhD Dissertation, University of California, San Diego.
- Carrow E (1974) A test using elicited imitations in assessing grammatical structure in children. *The Journal of Speech and Hearing Disorders*, **39**, 437–444.
- Chiat S (2015) Nonword repetition. In Armon-Lotem S, de Jong J and Meir N (eds), *Methods for assessing multilingual children: Disentangling bilingualism from language impairment*. Bristol: Multilingual Matters, pp. 125–150.
- Conti-Ramsden G, Botting N and Faragher B (2001) Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, **42**, 741–748.



- Dollaghan CA and Horner EA** (2011) Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research*, **54**, 1077–1088.
- Dollaghan C and Campbell TF** (1998) Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, **41**, 1136–1146.
- Eadie PA, Fey ME, Douglas JM and Parsons CL** (2002) Profiles of grammatical morphology and sentence imitation in children with specific language impairment and down syndrome. *Journal of Speech, Language, and Hearing Research*, **45**, 720.
- Ebert KD** (2014) Role of auditory non-verbal working memory in sentence repetition for bilingual children with primary language impairment. *International Journal of Language & Communication Disorders*, **49**, 631–636.
- Everitt A, Hannaford P and Conti-Ramsden G** (2013) Markers for persistent specific expressive language delay in 3–4-year-olds. *International Journal of Language & Communication Disorders*, **48**, 534–553.
- Fleckstein A, Prévost P, Tuller L, Sizaret E and Zebib R** (2016) How to identify SLI in bilingual children: A study on sentence repetition in French. *Language Acquisition*, **25**, 85–101.
- Gallimore R and Tharp RG** (1981) The interpretation of elicited sentence imitation in a standardized context. *Language Learning*, **31**, 369–392.
- Genesee F, Hamers J, Lambert WE, Mononen L, Seitz M and Starck R** (1978) Language processing in bilinguals. *Brain and Language*, **5**, 1–12.
- Gibson TA, Summers C, Peña ED, Bedore LM, Gillam RB and Bohman TM** (2015) The role of phonological structure and experience in bilingual children's nonword repetition performance. *Bilingualism: Language and Cognition*, **18**, 551–560.
- Gillam RB and Pearson NA** (2004) *Test of Narrative Language*. Austin: Pro-Ed.
- Gutiérrez-Clellen VF and Kreiter J** (2003) Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics*, **24**, 267–288.
- Gutiérrez-Clellen VF, Restrepo MA and Simón-Cerejido G** (2006) Evaluating the discriminant accuracy of a grammatical measure with Spanish-speaking children. *Journal of Speech, Language, and Hearing Research*, **49**, 1209–1223.
- Hair JF, Black WC, Babin BJ and Anderson RE** (2014) *Multivariate Data Analysis: Pearson New International Edition* (6<sup>th</sup> ed.). Upper Saddle River: Pearson Prentice Hall.
- Hesketh A and Conti-Ramsden G** (2013) Memory and language in middle childhood in individuals with a history of specific language impairment. *Plos One*, **8**, e56314.
- Hollingshead A.** (1975) *Four Factor Index of Social Status*. New Haven, CT: Yale University.
- Kapantzoglou M, Thompson MS, Gray S and Restrepo MA** (2016) Assessing measurement invariance for Spanish sentence repetition and morphology elicitation tasks. *Journal of Speech, Language, and Hearing Research*, **59**, 254–266.
- Kidd E, Brandt S, Lieven E and Tomasello M** (2007) Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and Cognitive Processes*, **22**, 860–897.
- Klem M, Melby-Lervåg M, Hagtvet B, Lyster S.-A. H, Gustafsson J.-E and Hulme C** (2015) Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, **18**, 146–154.
- Krishnan S, Alcock KJ, Mercure E, Leech R, Barker E, Karmiloff-Smith A and Dick F** (2013) Articulating novel words: Children's oromotor skills predict nonword repetition abilities. *Journal of Speech, Language, and Hearing Research*, **56**, 1800–1812.
- Leonard LB** (2014) *Children with Specific Language Impairment*. Cambridge: MIT Press.
- Lugo-Neris MJ, Peña ED, Bedore LM and Gillam RB** (2015) Utility of a language screening measure for predicting risk for language impairment in bilinguals. *American Journal of Speech-Language Pathology*, **24**, 426–437.
- McFarland J, Husser B, de Brey C, Snyder T, Wang X, Wilkinson-Flicker S, Gebrekristos S, Zhang J, Rathbun A, Barmer A, Bullock Mann F and Hinz S** (2017) *The Condition of Education 2017* (No. NCEES 2017-144). Washington, DC: National Center for Educational Statistics.
- McGee S** (2002) Simplifying likelihood ratios. *Journal of General Internal Medicine*, **17**, 647–650.
- Meir N** (2017) Effects of Specific Language Impairment (SLI) and bilingualism on verbal short-term memory. *Linguistic Approaches to Bilingualism*, **7**, 301–330.
- Meir N and Armon-Lotem S** (2017) Independent and combined effects of socioeconomic status (SES) and bilingualism on children's vocabulary and verbal short-term memory. *Frontiers in Psychology*, **8**, 1442.
- Meir N, Walters J and Armon-Lotem S** (2016) Disentangling SLI and bilingualism using sentence repetition tasks: the impact of L1 and L2 properties. *International Journal of Bilingualism*, **20**, 421–452.
- Melby-Lervåg M and Hulme C** (2010) Serial and free recall in children can be improved by training: evidence for the importance of phonological and semantic representations in immediate memory tasks. *Psychological Science*, **21**, 1694–1700.
- Moll K, Hulme C, Nag S and Snowling MJ** (2015) Sentence repetition as a marker of language skills in children with dyslexia. *Applied Psycholinguistics*, **36**, 203–221.
- Newcomer PL and Hammill DD** (2008) *Test of Language Development – Primary 4 (TOLD – P4)*. Austin: Pro-Ed.
- Oetting JB, McDonald JL, Seidel CM and Hegarty M** (2016) Sentence recall by children with SLI across two nonmainstream dialects of English. *Journal of Speech, Language, and Hearing Research*, **59**, 183–194.
- Peña ED, Bedore LM, Gutiérrez-Clellen VF, Iglesias A and Goldstein BA** (2010a) *Bilingual English Spanish Oral Screener (BESOS): Pre-K, 1st, and 3rd grade* (Unpublished test). Austin: University of Texas at Austin.
- Peña ED, Bedore LM, Lugo-Neris M and Albudoor N** (in press) Identifying developmental language disorder in school-age bilinguals: Semantics, grammar, and narratives. *Language Assessment Quarterly*.
- Peña ED, Bedore LM, Gutiérrez-Clellen VF, Iglesias A and Goldstein BA** (2010b) *Bilingual English-Spanish Assessment – Middle Extension, Experimental Version (BESA-ME)* (Unpublished test). Austin: University of Texas at Austin.
- Peña ED, Gutiérrez-Clellen VF, Iglesias A, Goldstein B and Bedore LM** (2018) *Bilingual English Spanish Assessment (BESA)*. Baltimore: Brookes.
- Plante E and Vance R** (1994) Selection of preschool language tests. *Language, Speech, and Hearing Services in Schools*, **25**, 15–24.
- Polišenská K, Chiat S and Roy P** (2015) Sentence repetition: What does the task measure? *International Journal of Language & Communication Disorders*, **50**, 106–118.
- Poll GH, Miller CA, Mainela-Arnold E, Adams KD, Misra M and Park JS** (2013) Effects of children's working memory capacity and processing speed on their sentence imitation performance. *International Journal of Language & Communication Disorders*, **48**, 329–342.
- Potter MC and Lombardi L** (1990) Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, **29**, 633–654.
- Redmond SM** (2005) Differentiating SLI from ADHD using children's sentence recall and production of past tense morphology. *Clinical Linguistics & Phonetics*, **19**, 109–127.
- Restrepo MA, Gorin JS, Gray S, Morgan GP and Barona N** (2010) Development of a language impairment screener for Spanish speaking children - SSLIC: Phase 1 - Task Development. Presented at the Society for Research on Educational Effectiveness, Washington, DC: SREE.
- Rice ME and Harris GT** (2005) Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, **29**, 615–620.
- Riches N** (2012) Sentence repetition in children with specific language impairment: An investigation of underlying mechanisms. *International Journal of Language & Communication Disorders*, **47**, 499–510.
- Riches N, Loucas T, Baird G, Charman T and Simonoff E** (2010) Sentence repetition in adolescents with specific language impairments and autism: An investigation of complex syntax. *International Journal of Language & Communication Disorders*, **45**, 47–60.
- Seeff-Gabriel B, Chiat S and Dodd B** (2010) Sentence imitation as a tool in identifying expressive morphosyntactic difficulties in children with severe speech difficulties. *International Journal of Language & Communication Disorders*, **45**, 691–702.
- Semel E, Wiig EH and Secord WA** (1995) *Clinical Evaluation of Language Fundamentals, Third Edition*. San Antonio: The Psychological Corporation, Harcourt Brace & Co.

- Simón-Cerejido G** (2017) Sentence repetition in typical and atypical Spanish-speaking preschoolers who are English Language Learners. In Auza Benavides A and Schwartz RG (eds), *Language Development and Disorders in Spanish-speaking Children*. Cham: Springer International Publishing, pp. 205–215.
- Simon-Cerejido G and Méndez LI** (2018) Using language-specific and bilingual measures to explore lexical-grammatical links in young Latino dual-language learners. *Language, Speech, and Hearing Services in Schools*, **49**, 537–550.
- Smolík F and Vávru P** (2014) Sentence imitation as a marker of SLI in Czech: Disproportionate impairment of verbs and clitics. *Journal of Speech, Language, and Hearing Research*, **57**, 837–849.
- Stark RE and Tallal P** (1981) Selection of children with specific language deficits. *Journal of Speech and Hearing Disorders*, **46**, 114–122.
- Stokes SF, Wong A. M.-Y, Fletcher P and Leonard LB** (2006) Nonword repetition and sentence repetition as clinical markers of specific language impairment: The case of Cantonese. *Journal of Speech, Language, and Hearing Research*, **49**, 219–236.
- Summers C, Bohman TM, Gillam RB, Peña ED and Bedore LM** (2010) Bilingual performance on nonword repetition in Spanish and English. *International Journal of Language & Communication Disorders*, **45**, 480–493.
- Thordardottir E and Brandeker M** (2013) The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders*, **46**, 1–16.
- Tomblin JB, Records NL, Buckwalter P, Zhang X, Smith E and O'Brien M** (1997) Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research*, **40**, 1245–1260.
- Topbaş S, Kaçar-Kütükçü D and Kopkalli-Yavuz H** (2014) Performance of children on the Turkish Nonword Repetition Test: Effect of word similarity, word length, and scoring. *Clinical Linguistics & Phonetics*, **28**, 602–616.
- Tuller L, Hamann C, Chilla S, Ferré S, Morin E, Prevost P, Dos Santos C, Abed Ibrahim L and Zebib R** (2018) Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders*, **53**, 888–904.
- Verhoeven L, Steenge J, van Weerdenburg M and van Balkom H** (2011) Assessment of second language proficiency in bilingual children with specific language impairment: a clinical perspective. *Research in Developmental Disabilities*, **32**, 1798–1807.
- Willis CS and Gathercole SE** (2001) Phonological short-term memory contributions to sentence processing in young children. *Memory*, **9**, 349–363.
- Ziethe A, Eysholdt U and Doellinger M** (2013) Sentence repetition and digit span: Potential markers of bilingual children with suspected SLI?. *Logopedics, Phoniatrics, Vocology*, **38**, 1–10.